Geo-Summaries based on CRF

- Geographic Summary
- Minorthird
- Tests with HAREM's Golden Collections
- Geo-Net-PT
- HENDRIX
- Results with 1,5 Gbytes from WPT05
- TODO

Geo-Summaries based on CRF

Geographic Summary

Input: document

Output:

- minimal list of resolved geographic references
- RDF/XML

Geo-Summaries based on CRF

- **Minorthird** supervised machine learning software to extract entities and categorize text. It was written primarily by William W. Cohen, professor at Carnegie Mellon University
- Contains methods to visualize both training data and the performance of classifiers
- Complex hand-coded annotation programs can be implemented with a special-purpose annotation language called Mixup
- Handles input format as ”XML” (no overlappings)

Geo-Summaries based on CRF

"Fernão de Magalhães morre nas **<local>Filipinas</local>** no curso daquela expedição, posteriormente chefiada por Juan Sebastián Elcano 1522. Alista-se com 25 anos na armada que foi à **<local>Índia</local>** comandada por Francisco de Almeida, embora o seu nome não figure nas crónicas; sabe-se no entanto que ali permaneceu oito anos, que esteve em **<local>Goa</local>**, **<local>Cochim</local>**, **<local>Quíloa</local>**, que acompanhou Diogo Lopes de Sequeira a **<local>Malaca</local>**, viagem que acabou em naufrágio."

Geo-Summaries based on CRF

Training with HAREM's Golden Collections

	MiniHAREM	HAREM I	HAREM II
Size	514 Kbytes	734 Kbytes	1.1 Mbytes
Unique PLACE entities	397	514	612
Total PLACE entities	792	1146	1200

Train:
MiniHAREM
HAREM I

Test:
HAREM II

Precision: 64%
Recall: 45%
F-Measure: 53%

Geo-Summaries based on CRF

- **Some tags don't include feature types, ex:**

no **concelho de** <EM ID="2284" CATEG="LOCAL" TIPO="HUMANO">Bardez</EM>

o mercado do **distrito de** <EM ID="843" CATEG="LOCAL" TIPO="HUMANO">Santarém</EM>

- **Some entities can generate garbage on the learning phase:**

”Estação <EM ID="1402" CATEG="LOCAL"TIPO="HUMANO">Abesses</EM> Ônibus Linhas <EM ID="1403" CATEG="LOCAL" TIPO="HUMANO">30</EM> <EM ID="1404" CATEG="LOCAL" TIPO="HUMANO">54</EM> <EM ID="1407" CATEG="LOCAL" TIPO="HUMANO">74</EM> <EM ID="1408" CATEG="LOCAL" TIPO="HUMANO">81</EM>”

- **Suggestion: re-annotation of HAREM's GC regarding geographic information extraction**

Geo-Summaries based on CRF

Geo-NET-PT-01

- Adminstrative data from Portugal
- Províncias, Distritos, Concelhos, Ilhas, Munícipios, Freguesias, Zonas, Localidades, Arruamentos(rua, largo, praça, estrada, etc)
- Relationships among features, mainly part-of

Geo-NET-PT-02

- Data from Geo-NET-PT01 + physical and geographical data from Portugal

Geo-Summaries based on CRF

HENDRIX is an **E**ntity **N**ame **D**esambiguator and **R**ecognizer for **I**nformation e**X**traction

- Handles extracted entities
- Looks for matches in Geo-Net-PT
- Use of caches for found and not found entities → reduces number of queries, increases (alot!) speed.
- For a given document stores:

  resolved/not resolved entities;

  term frequency;

  positions in the document;

  possible meanings from Geo-Net-PT;

Geo-Summaries based on CRF

- Use of regular expressions for handling feature types:

  extracted term: 'Avenidade da Liberdade'

  queries Geo-Net-PT to give all features whose name is 'Liberdade' and type 'Avenidade'

  #features named 'Liberdade' → **486**

  #features named 'Liberdade' and feature type 'AVE' → **69**

- naïve geo-summaries, discards geo-entities that do not have any relation with any other resolved geo-entitiy in the document

Geo-Summaries based on CRF

Extraction from 1,5 Gbytes of WPT-05

1st: Language Classification with NgramJ

2nd: HADOOP cluster set-up with Minorthird

3rd: HENDRIX processes extracted entities (portuguese)

Sets of Entities	Unique Entities	Percentage
Resolved	2 718	11.71%
Resolved holding Feature Type	529	2.28%
Unresolved	18 823	81.08%
Unresolved holding Feature Type	1144	4.93%
Total	23 214	100%

Geo-Summaries based on CRF

Entities out of Geo-Net-PT scope

Entity	Number of occurrences	Entity	Number of occurrences
Iraque	1576	Bósnia	199
Los Angeles	965	América do Sul	183
Kosovo	881	Costa Rica	175
Europa Central	785	São Leopoldo	162
Bagdad	523	Avenida Paulista	34
Nova Iorque	431	Parque do Ibirapuera	9
América Latina	299	Avenida Recife	4

TODO: use an World Ontology to analyze how many extracted entities are outside Geo-Net-PT's scope

Geo-Summaries based on CRF

Some interesting extracted entities:

Aeroporto de Lisboa 2
Aeroporto Sá Carneiro 2
Alfândega do Porto 7
Alto Douro 2
Baixa do Porto 86
Barragem do Sabor 2
Camara Municipal do Porto 3
Capela dos Ossos 2
Casa da Música 18
Casino da Póvoa 2
Casino de Vilamoura 2
Coliseu do Porto 26
Coliseu dos Recreios 3
Convento das Freiras 2
Convento de Alpendorada 2
Estádio da Luz 7
Estádio das Antas 6
Estádio de Alvalade 2
Estádio do Algarve 8
Estádio do Dragão 10
Hospital de S. João 26
Hospital São Sebastião 8
Igreja de Monte Abrão 2
Igreja de São João 2
Mercado do Porto 1
Mercado Ferreira Borges 2
Museu Soares do Reis 3
Palácio de Belém 3
Palácio de Cristal 1
Palácio de S. Bento 2
Pavilhão Atlântico 7
Praia da Carrapateira 2
Praia da Galé 2
Praia das Maças 1
Praia de Carcavelos 135
Praia do Magoito 2
Praia do Norte da Nazaré 2
Serra de Malcata 1
Serra de Nogueira 2
Serra de São Pedro 2
Serra de Silves 8
Serra dos Candeeiros 2
Serra Estrela 1
Serra Gardunha 1
Teatro Nacional de São João 2
Teatro Sá da Bandeira 5
Teatro São Luiz 2
Teatro Tivoli 2
Teatro Villarett 1
Torre das Antas 9
Torre do Tombo 17
Torre dos Clerigos 1
Torre Eiffel 1
Jardim da Amália 2
Jardim da Cordoaria 2
Jardim da Estrela 3
Jardim da Gulbenkian 2
Jardim de Oeiras 2
Jardim do Bonfim 1
.....
.....
...
..
..
.
..

Geo-Summaries based on CRF

Implicit geographic references

Entity Type	N° Unique Entities	Total
Airport	9	20
Dum	3	4
Chapel	5	5
Church	6	10
Covent	3	6
Forum	4	9
Stadium	11	32
Museum	9	20
Markets	3	10
Palace	4	10

Entity Type	N° Unique Entities	Total
Beach	45	221
Sierra	39	137
Theater	15	39
Coliseum	2	29
Casino	3	5
Vila	14	24
Garden	12	20
Parks	76	231
Historical Buildings	2	18
Total	265	850

Geo-Summaries based on CRF

TODO

- Use semantic similarity to disambiguate geographic references;
- Do entity extraction on the portuguese pages from WPT05 (~ 25 Gigabytes)
- Same thing for Wikipedia articles with Geo-Coordenates (for evaluation)
- Finish thesis
- Re-anotate Golden Collections
- Explore in more detail M3rd/Mixup for generating new features in the learning phase